Numerical Methods

Dr. Phonindra Nath Das

Department of Mathematics
Ramakrishna Mission Vivekananda Centenary College

Email: *phonimath@gmail.com*

September 3, 2021

Outline of presentation

Introduction

- In numerical analysis, we deals with numbers like rational or irrational, real or complex. Here we talks about real numbers.
- Out of infinitely many digits in a real numbers a computational device can hold only a finite number of digits.
- So, after a finite number of digits, depending on the capacity of the computational device, the rest should be discarded in some sense.
- In this way, the representation of the real number on a computing device is only approximate. Although, the omitted part of that number is very small in its value, this approximation can lead to considerably large error in the numerical computation. Here, we study error due to approximating numbers.
- We also study the various operators while handling certain given data set or numbers.

Approximate numbers

In numerical calculations two types of numbers are used namely exact and approximate.

Numbers in which there is no approximation or ambiguity present are called exact

For example 2, $\frac{1}{5}$, $2\frac{1}{3}$, 3^4 etc are exact numbers. On the other hands numbers like $\sqrt{3}$, e, π are all exact but can not be represents as finite number of digit, while in numerical calculations we often represents $\sqrt{3}$ as 1.732, π as 3.142. Therefore, 1.732 and 3.142 are called approximate numbers of $\sqrt{3}$ and π.

Numbers which are approximations to exact ones and differ by very tiny quantity are known as approximate numbers

Significant Figures

In numerical calculation significant figures is very important.

Significant figures

While representing a number the digits which are used known as **significant figures**. Some time it is called significant digits also.

Suppose we have the numbers 1.7634, 4.08721, 0.2019, now clearly these numbers contains 5, 6 and 4 digits respectively. And hence they are respectively called the numbers of 5, 6 and 4 significant figures.

Now look at the number 0.0435, it contains 4 places of decimal but it is a 3 significant figures number as the leading zero used to fix the position of the decimal place.

Thus 0.0082, 420, 8.0410 and 21.653 have 2, 3, 4 and 5 significant figures respectively.

Rounding off Numbers

Let a number be x and a decimal representation of x is generally given by

$$x = a_p \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-q}$$

which is also represented as

$$a_0 + a_1 \times 10 + a_2 \times 10^2 + \cdots + a_p \times 10^p + a_{-1} \times 10^{-1} + a_{-2} \times 10^{-2} + \cdots + a_{-q} \times 10^{-q}$$

where a_i's takes one of the values of $0, 1, \cdots, 9$, $a_p \neq 0$ and some times $q \to \infty$ (q goes to infinite in case of irrational number).

Now the rejection of unwanted digits of a given number is known as **rounding off** of that number

Rules of rounding off a number

Let the given number be

$$x = a_p \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-q}$$

and we propose to retain digits up to n places of decimal where $n < q$ in general.

i) Discard all digits to the right of the nth place and if the discarded number is less than half a unit in the nth place i.e., they are one of 0, 1, 2, 3 and 4, then leave the nth place unchanged and the rounded off number will be

$$x = a_p \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-n}$$

Example: Let 2.37564829 to be rounded off correct up to 5 significant figure. Here discarding all digits right after 6 we get 2.3756 and the discarded figure is 4 which is less than half of unit, hence rounded off correct up to 5 significant figure of the given number will be 2.3756

Rules of rounding...

ii) If the discarded number is greater than half a unit in the nth place, then add 1 to the nth digit and the rounded off number will be

$$x = a_p \cdots a_1 a_0 a_{-1} a_{-2} \cdots (a_{-n} + 1)$$

Example: Let 2.37564829 to be rounded off correct up to 6 significant figure. Here discarding all digits right after 4, we get 2.37564 and the discarded figure is 8 which is greater than half of unit, hence rounded off correct up to 6 significant figure of the given number will be 2.37565 (note that we have added 1 to the 6-th digit).

Rules of rounding...

iii) If the discarded number is exactly half a unit in the nth place, then leave the nth place unchanged if it is an even number and the rounded off number will be

$$x = a_p \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-n}$$

Example: Let 1.3574529 to be rounded off correct up to 5 significant figure. Here the 6th figure that is the discarded digit is exactly equal to half (which is 5) and the 5th significant figure is 4 which is even, so we keep fifth significant figure 4 unaltered, hence the rounded off number up to 5th significant figure in this case is 1.3574.

Rules of rounding... & significant figures

iv) If the discarded number is exactly half a unit in the nth place, then add 1 to the nth digit if it is an odd number and the rounded off number will be

$$x = a_p \cdots a_1 a_0 . a_{-1} a_{-2} \cdots (a_{-n} + 1)$$

Example: Let 3.2527529 to be rounded off correct up to 5 significant figure. In this case the 6th figure that is the discarded digit is exactly equal to half (i.e., the digit is 5) and the 5th significant figure is 7 which is odd, so we add 1 to the fifth significant figure, hence the rounded off number up to 5th significant figure in this case is 3.2528.

Significant figures

A number is said to be correct up to n-significant figures, if the number be rounded off to n-figures according to the above laws

Errors

When we approximate certain number for numerical calculations we must drop some significant figures which is basically error. We can avoid this types of error for smoothness of numerical analysis, so we need to understand about various errors.

Absolute error

The absolute error E_a of a number is defined as modulus the difference between its true value V_t and its approximate value V_a. So, the absolute error $E_a = |V_t - V_a|$

Relative error

The relative error E_r of a number is known as the absolute error divided by its true value. Therefore, the relative error $E_r = \frac{E_a}{V_t} = \frac{|V_t - V_a|}{V_t}$

Errors...

Relative percentage error

The relative percentage error E_p is defined as the relative error multiplied by 100, then the relative percentage error $E_p = E_r \times 100 = \frac{E_a}{V_t} \times 100 = \frac{|V_t - V_a|}{V_t} \times 100$

Remark:

If a number is rounded to m decimal places then the absolute error satisfies $E_a \leq \frac{1}{2} \cdot 10^{-m}$

$0.00028 < 0.0005$ $2.05672 \cong 2.057$ $E_a \leq \frac{1}{2000}$

Theorem on rounding off and relative error

Theorem

If a number be rounded off to n significant figures, then the relative error $E_r < \frac{1}{p \times 10^{n-1}}$ where p is the first significant figure in the number and $n > 1$

Proof: Let V_t is the true value and m be the number of decimal places in approximate value V_a correct up to n significant figures.
When $m < n$, then V_t has $n - m$ digits in the integral part and we have $E_a \leq \frac{1}{2} \cdot 10^{-m}$.

$$\therefore |V_t| \geq p \times 10^{n-m-1} - \frac{1}{2} \cdot 10^{-m} = \frac{1}{2} \cdot 10^{-m}(2p \times 10^{n-1} - 1)$$

$$\text{so, } E_r = \frac{E_a}{|V_t|} \leq \frac{\frac{1}{2} \cdot 10^{-m}}{\frac{1}{2} \cdot 10^{-m}(2p \times 10^{n-1} - 1)} \leq \frac{1}{2p \times 10^{n-1} - 1}.$$

Proof continuing...

Since n and p are two positive integers such that $1 \leq p \leq n$ and $n > 1$, then we have

$$\begin{aligned} & 2p \times 10^{n-1} - 1 > p \times 10^{n-1} \\ \Longrightarrow \quad & E_r \leq \frac{1}{2p \times 10^{n-1} - 1} < \frac{1}{p \times 10^{n-1}} \end{aligned}$$

If $m = n$, V_t has no digit in the integral part, hence V_t is purely decimal number with p as its first decimal figure. Thus,

$$\begin{aligned} E_a &\leq \frac{1}{2} \cdot 10^{-m},\ |V_t| \geq p \times 10^{-1} - \frac{1}{2} \cdot 10^{-m} \\ &= \frac{1}{2} \cdot 10^{-m}(2p \times 10^{m-1} - 1) \\ \therefore\ E_r &= \frac{E_a}{|V_t|} \leq \frac{\frac{1}{2} \cdot 10^{-m}}{\frac{1}{2} \cdot 10^{-m}(2p \times 10^{m-1} - 1)} = \frac{1}{2p \times 10^{m-1} - 1}. \end{aligned}$$

Proof continuing...

$$\implies E_r \leq \frac{1}{2p \times 10^{n-1} - 1} \quad (\because\ m = n)$$
$$< \frac{1}{p \times 10^{n-1}}$$

Finally when $m > n$, the first significant figure appears in the $(m - n + 1)$th place and thus, we have $E_a \leq \frac{1}{2} \cdot 10^{-m}$ and

$$|V_t| \geq p \times 10^{n-m-1} - \frac{1}{2} \cdot 10^{-m} = \frac{1}{2} \cdot 10^{-m}(2p \times 10^{n-1} - 1)$$
$$\therefore\ E_r = \frac{E_a}{|V_t|} \leq \frac{\frac{1}{2} \cdot 10^{-m}}{\frac{1}{2} \cdot 10^{-m}(2p \times 10^{n-1} - 1)} \leq \frac{1}{2p \times 10^{n-1} - 1}$$
$$< \frac{1}{p \times 10^{n-1}}.$$

Hence the theorem.

Concluding remarks!

You are advised to follow this slides again and read any text book you have.

In the next class we will study about **Finite Difference**